MOVIE SALES PREDICTION

DENG, Ken
20400660
kdengab@connect.ust.hk

HOU, Jiefeng
20361723
jhouad@connect.ust.hk

LAU, Kim Kwan
10254702
kklauab@connect.ust.hk

WONG, Lok Heung
20323343
lhwongaj@connect.ust.hk

ABSTRACT

Film studios collectively produce several hundred movies every year. The budget of these movies is of the order of hundreds of millions of dollars. Therefore, making their box office successful by knowing which movies are likely to succeed and which are likely to fail before the release is absolutely essential for the survival in the industry. Also, conducting prediction could help them to determine the most appropriate release date for a movie by looking into the overall market, so the prediction of the sales of movies is crucial to the industry. Machine learning algorithms are widely used to make predictions such as growth in the stock market, demand for products and nature of tumors. This paper presents a detailed study using Linear Regression and Random Forest to predict two weeks sales for upcoming movies.

Keywords

Data mining, Linear Regression, Random Forest, Machine Learning

1. INTRODUCTION

In China, the film industry is growing rapidly. The box office income increased from 1.6 billion to 21.769 billion from 2005 to 2013. If the population of China is 1.29 billion, the average ticket fee is $ 35, and the annual income in 2013 is only $ 21.7 billion, indicating that the number of Chinese watching movies is less than half of the total population. As the saturated market rate of United States is 4.3, China still has a lot of space for development. In fact, the box office growth rate is 27.5% each year. It is estimated that if China is not suffering from economic downturn, the growth rate is promising until 2020.

This shows that film industry will become one of the most potential markets in China. The movie industry is in dire need of software to predict the movie income. In this study, we attempt to use the Chinese market information of movie from 2012 - 2016 to predict the gross revenue of the movies[1].

1.1 Python: scikit-learn

In this study, we mainly use scikit-learn as the development tool. Scikit-learn is a free software machine learning library for Python programming language[4]. It features various classification, regression and clustering algorithms including support vector machines (SVM), random forests, gradient boosting, k-means and DBSCAN. We will use linear regression and random forests in our project.

1.2 Roadmap

Section 2 describes the role of dataset collection and preprocessing in data mining. Feature Engineering and Selection have discussed in section 3. Algorithms and Results are shown in section 4 and section 5 respectively. Section 6 concludes the paper. Finally, improvement is suggested in section 7 and future work of title is in section 8.

General Design is shown in the Figure 1.

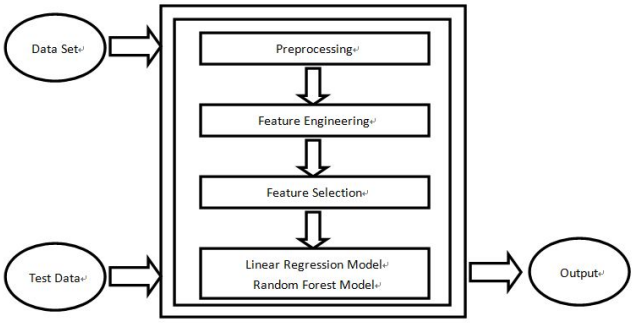

Figure 1: General Design

2. Data Collection and Preprocessing

Data Collection

The initial dataset to be used for movie sales prediction mainly includes four CSV files: metadata.csv, relation.csv, box.csv, score.csv. Metadata file includes four attribute: movie name, type, release time and total box. Box file includes seven attributes: movie name, release time, current time, daily box, total box, daily screen, daily screen percent. Relation file includes server multi-class attribute which will be clearly described below in this chapter. The score file will be included during prediction because the rating of the movie will not be released before the movie releases. Additionally, the rating of movie does not have much effect on box-office. The details of relation file are shown below:

attribute	type	Description
actor	nominal	All the movies include variable number of actors (3~50). Most of them are in

		Chinese.
director	nominal	All the movies include one or two directors. Most of them are in Chinese.
publisher	nominal	Since the quantity of publisher is small, all publishers are regarded as producer.
producers	nominal	Producer is also a multi-label attribute, and producers are both in Chinese and English.

Table 1: Details of Attributes

The attributes above are the main features in the prediction model. However, all of them are nominal attribute, which need to be normalized. Another difficulty is multi-label problem. Therefore, it is necessary to collect external data online. According to the names of actors and directors, to search the movie related to them will show a clear rating of actors or director. In this way, some numerical features can be setup for linear model or linear regression model.

Since all the movie data is China box-office, the information on twitter or IMDB does not have reference value. Our target website is DOUBAN(https://www.douban.com/), one of the most popular social media platforms in China. DOUBAN has tremendous amount of comments about movies shown in China. Additionally, DOUBAN provides the APIs to public for free, but this APIs are limited to only request 100 times in a hour. The web GET method is https://api.douban.com/v2/movie/search?tag=keyword. The format of the web APIS is json, the example is as below:

{"count": 20,

"start": 0,

"total": 200

…

"collect_count": 794708…}

In this case, we build a web application on .NET platform. To search actor or director name as parameter, DOUBAN APIs will respond with all the movie information related to it. We mainly pick up two attributes as potential features:

1. Total, this attribute stands for the amount of actor or director's works in the past. This number can be a symbol of the strength.

2. Collect_count, this attribute stands for the amount of subscribers or followers of a particular movie. This numerical attribute shows the popularity of the movie.

For actor, we compute average of collect_count => Actor_rate1. We compute the average of Total => Actor_rate2.

For director, a single movie usually has less than two director, so we compute the average of collect_count of them => Director _rate.

attribute	type	range
Actor_rate1	numerical	{0~5000000}
Actor_rate2	numerical	{0~200}
Director_rate	numerical	{0~5000000}

Table 2: Range of Attributes

Data Visualization

According to the result of the basic analysis of given data, it is easy to find out the attribute month has a positive relationship with movie's sales. The below graphs show that the monthly average sales and total sales separately from 2012-2016.

Figure 2: Monthly average sales of each movie

Figure 3: Monthly movie's total sale

These graphs show that the top three months are February, June and July, because the Spring Festival is in February and school summer holiday is from June to August. Therefore, the holiday may be closely related to movie's sales from the basic analysis.

It is difficult to identify if the movie release date is within Chinese holidays or not, so using a web spider with Chinese holiday API (http://www.easybots.cn/api/holiday.php?d=20130101) to get the results (working day = 0, weekends = 1, holiday = 2) is a simple and effective method.

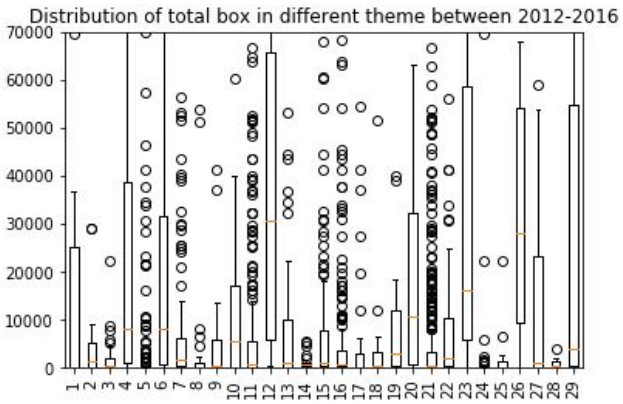

Figure 4: Distribution of total box in different theme between 2012 - 2016

{1: '亲情', 2: '传记', 3: '儿童', 4: '冒险', 5: '剧情', 6: '动作', 7: '动画', 8: '励志', 9: '历史', 10: '古装', 11: '喜剧', 12: '奇幻', 13: '家庭', 14: '恐怖', 15: '悬疑', 16: '惊悚', 17: '战争', 18: '歌舞', 19: '武侠', 20: '灾难', 21: '爱情', 22: '犯罪', 23: '科幻', 24: '纪实', 25: '纪录片', 26: '警匪', 27: '青春', 28: '音乐', 29: '魔幻'}

From the above box plot, we can found that science fiction, fantasy and gangster film may attract more people to watch.

A. Data Preprocessing

After the data collection, we removed the data which is meaningless to the prediction. We removed the data which does not have first 14 days daily box or 14 days daily box. Also, data with invalid character (such as '[]') is removed. Data which does not have the movie's name is deleted.

B. Data Integration

In Figure 5, the upper part is the original data and the bottom part is the external data retrieved from DOUBAN and CBOOO. In this step, integrated data are transformed or consolidated, so that the process can be more efficient.

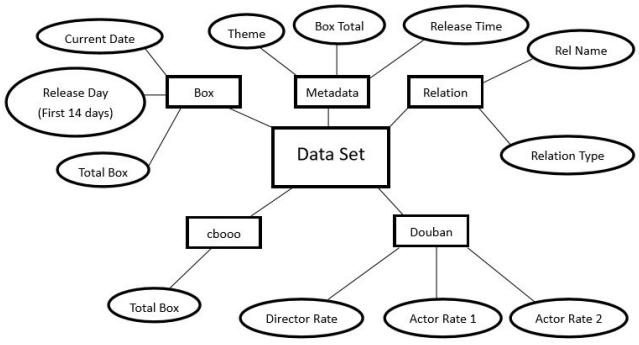

Figure 5: Diagram of Data Integration

C. Data Transformation

After integrating the data, the dataset is mixed with both nominal and numerical attributes, for a linear process or regression process, only the numerical features can be added to the model. Therefore, the first step is to rescale data. Many machine learning algorithms can benefit from rescaling the attributes to all have the same scale. Often this is referred to as normalization and attributes are often rescaled into the range between 0 and 1. We use MinMaxScaler class from scikit-learn to complete this step:

scaler = MinMaxScaler(feature_range=(0, 1))

rescaled = scaler.fit_transform()

Standardization is useful for transforming attributes with a Gaussian distribution and differing means and standard deviations to a standard Gaussian distribution with a mean of 0 and a standard deviation of 1. Therefore, the second step is to standardize the rescaled data. It is suitable for techniques that assume a Gaussian distribution in the input variables and work better with rescaled data, such as linear regression, logistic regression and linear discriminate analysis. We use StandardScaler from scikit-learn to complete this step:

scaler = Normalizer().fit()

normalized = scaler.transform()

nominal features like 'year', 'month', 'theme' and 'producer' are prepared for the random forest model, we turn them into label number in the Data Integration process. For example, themes like 'animation' will turn to 24, a unique number(ID).

3. Feature (Engineering & Selection)

By the use of Pearson correlation coefficient, we can understand a feature's relation to the response variable especially to measure linear correlation between two variables. Since only directorrate, actor rate1, actor rate 2 and totalBox14 are continuous numerical variables , we use those variables to measure correlation.

	directorrate	actor rate1	actor rate2	totalBox14
directorrate	1.000	0.137	0.271	0.235
actor rate1	0.137	1.000	0.553	0.087
actor rate2	0.271	0.553	1.000	0.357
totalBox14	0.235	0.087	0.357	1.000

Table 3: Correlation between features

From the above figures, we can conclude that actor rate 2 has the highest correlation to totalBox14. It means that the amount of the past movie of the actor participant will affect the total income of the first 14 days obviously.

Having too many irrelevant features in the dataset can decrease the accuracy of the movie prediction model. Thus, in order to improve the accuracy of movie prediction model, it is necessary to perform features selection. Finally, a features selection method called "feature importance ranking" (provided by the scikit-learn Python library) has been selected. Specifically, this method can compute the relative importance of each attribute, the higher value means that the feature is more useful. Below table displays the relative feature importance values of movie dataset.

Feature	Type	Importance Value
director rate	float	0.10006
actor rate 1	float	0.10398
actor rate 2	float	0.10802
theme 1	category	0.09406
theme 2	category	0.5579
theme 3	category	0.03619
year	category	0.06633
month	category	0.09314
day	category	0.10953
product 1	category	0.09376
product 2	category	0.06704
product 3	category	0.04078
isHoliday	category	0.03132

Table 4:The value of feature importance

From the table 4, it is clear that the feature “director rate”, “actor rate 1”, “actor rate 2”, “theme 1”, “month”, “day”, “product 1” have a higher importance value than the others. Therefore, these features may will be selected.

4. Algorithms

In order to evaluate different models, we will use mean square error (MSE) as measure of the quality of an estimator[6]:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$$

If the hat of Y is a vector of n predictions, and Y is the vector of observed values related to the inputs to the function which generated the predictions.

For readability, we also use root mean square error (RMSE) as our measure:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2}$$

A. Baseline model by Zero Rule Algorithm

Baseline is necessary to be a measure of different algorithm. Therefore, the mean of the 14 days of sales observed in the training data was used to calculate MSE[7].

Implementation

1. Using cross validation randomly to split the entire data set into a training set and a test set, in the ratio of 80% to 20%.
2. Calculate the mean of the total box 14 of training data set
3. Use the mean as the result of prediction of testing data to calculate MSE and RMSE

B. Multiple Linear Regression

Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data[8].(Linear Regression is provided by scikit-learn)

Implementation

1. Using cross validation randomly to split the entire data set into a training set and a test set, in the ratio of 80% to 20%.
2. Linear regression is better to use continuous numerical variables, so director rate, actor rate1 and actor rate2 will be selected to fit the model.
3. In order to prevent overfit, we will calculate different MSE and RMSE by using different combination of features.

C. Random Forests

The Random Forests algorithm is one of the best among all classification algorithms, as it enables the classification of a large amount of data with accuracy. Also, the random forest can handle binary features, categorical features and numerical features[3][5].

Implementation

1. Using cross validation randomly to split the entire data set into a training set and a test set, in the ratio of 80% to 20%.
2. In order to handle categorical variables, OneHotEncoder (which provided by the scikit-learn) is used to transform categorical features to numeric feature.
3. Using RandomForestClassifier (which provided by the scikit-learn) with a parameter (n_estimators, which the number of trees in the forest, more trees reduces overfitting but takes longer to run) to predict.
4. Refit the random forest to the entire training set, using different n_estimators value to get the result.

5. Results

A. Baseline Model by Zero Rule Algorithm (correct to 3 decimal places)

The mean of the total box 14 of training data set	MSE (Mean Square Error)	RMSE(Root of Mean Square Error)
6286.724	254553048.700	15954.719

Table 5: Result of Zero Rule Algorithm

From the result, we expect that the later models using features will get a lower MSE than the above. This is our baseline to evaluate the performance of our models.

B. Multiple Linear Regression (correct to 3 decimal places)

Feature used	MSE (mean square error)	RMSE (root of mean square error)
director rate	291422207.322	17071.093
actor rate1	253213009.408	15912.668
actor rate2	202985289.109	14247.291
director rate , actor rate1	289276204.517	17008.122
director rate , actor rate2	231276499.236	15207.778
director rate , actor rate1, actor rate2	220727988.273	14856.917

Table 6: Result of Linear Regression

From the above result, we can found that actor rate 2 has the lowest MSE compared to other features' combination.

C. Random Forest (correct to 3 decimal places)

Using the average of 100 times MSE and RMSE results with same parameter.

n_estimators	MSE (mean square error)	RMSE (root of mean square error)
10	204948986.459	14316.039
20	219511510.631	14815.921
30	178162962.892	13347.770
40	153312670.243	12381.949
50	141656736.360	11901.963
100	96063386.6396	9801.193

From the above table, we can get a best result when n_estimators = 100 (tree number is 100).

D. Comparison of different algorithms (correct to 3 decimal places)

Method	Baseline model by Zero rule algorithms	Linear Regression(using actor rate2 as feature only)	Random Forest Decision (n_estimators = 100)
MSE (Mean Square Error)	254553048.754	202985289.109	96063386.6396
RMSE(Root of Mean Square Error)	15954.719	14247.291	9801.193

Table 7: Result of Algorithms

From the above result comparison, we can conclude that Random Forest can get a better result with the lowest MSE value. However, although we get a better result using Random Forest Decision, but the prediction performance is not well (since the error is high). We will analyze the reasons and try to improve in next section.

6. Conclusion

With the growing market in Chinese filming industry, the prediction of the sales of upcoming movies has become more important. The focus for this paper is using different features of movies to build a suitable model for the prediction of the total box office of upcoming movies within two weeks.

According to the results of mean square error (MSE) in the three models, random forest model represents the best prediction results. However, all predictions are still far from the real data. The main reason is the low correlation between features and sales. Although we worked hard on the data preprocessing, the results are still not satisfying. Also, most of the attributes of raw data are

categorical and multi-labelled. If we choose decision tree algorithm to select features, we should pick more categorical features and split them into the continuous features. Finally, the dimension and quality of the features are low. Therefore, using filters or wrappers to reduce the number of features could not improve the accuracy of results.

More explanations for the suitability of features will be discussed in the following.

One reason for the inaccurate prediction is that the raw attributes ,such as 'actor' and 'director',do not affect the sales in the box office heavily as expected. In other words, these raw attributes are relatively less important for prediction. Besides, China national policies have influence on the monthly sales in box office. For example, as July and August are 'Domestic protection month', foreign movies cannot be released in these two months. Therefore, unusual high sales are recorded for local cartoon movies or other dark horses. Another example is that during Spring festival, a long public holiday in China, no foreign movies are allowed to release in China and the box office will solely be occupied by 'New Year Movie'.

7. Improvement

Since all the raw data are string format, if we expect to build a better prediction model, there are two methods:

1. Random forest algorithm

Random forest model needs more accurate categorical features. Therefore, 'actor', 'producer', 'director' should be the most important attribute which are all multi-label. To select the most valuable elements can affect the accuracy of the random forest model directly.

2. Linear regression algorithm

Collecting more external data from social media platform can show rating of a particular movie. some external data we expect to collect in future is as below:

Promotion events of the movie

Type of movie (2D, 3D)

Country of movie

Producer rating

So far, we only compute average or standard score as the final rating of single attribute like 'actor'. While external data collection is necessary, to figure out an appropriate algorithm to compute accurate rating for particular feature is also important.

8. Future Work

Last chapter elaborated the works about the improvement of the raw data. However, collecting external data solely on the basis of the raw attributes and improving the algorithm are far from sufficient. The subjective attributes including ratings and comments, do not perform well in both random forest algorithm and linear regression algorithm. In this chapter, we will introduce a potential independent attribute: 'search volume'. Remarkably, although the search volume will reach the peak during the release period, we cannot predict the sales after the movie has already released. Therefore, the most valuable data is the search volume one week before the movie releases[2].

BAIDU INDEX, another Google trends in China, will be the target data source, because Google has been forbidden in China. Additionally, our target is to predict the movie box office in China, so the data from BAIDU is much more valuable than any other web search engines or social media like Google, Twitter, IMDB. Here is a trend chart example for movie 'Iron Man 3':

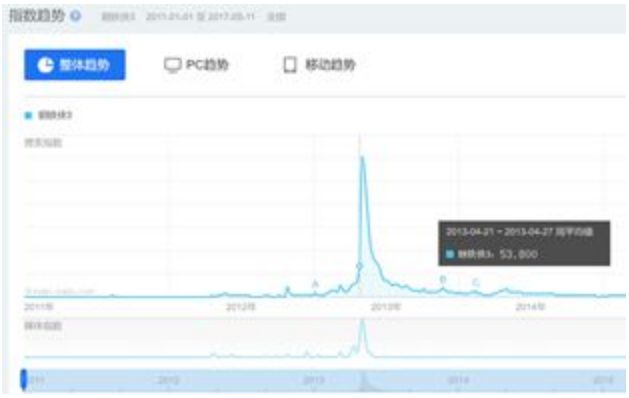

Figure 6: DAIDU INDEX

The chart above is collected from BAIDU INDEX. It shows the search volume trend. Obviously, the peak of the trend is in the release period and the exact value of search volume a week before released date is 53800. We also collect some other movies' search volume trend as the table:

movie name	release date	search volume	totalbox
钢铁侠3	2013-05-01	53800	61936.5
百星酒店	2013-02-01	6233	4828.82
辛巴达历险记	2013-05-31	4954	3539.4
大上海	2012-12-21	14653	10399.08
没女神探	2015-12-15	1215	99.9
公路美人	2015-04-03	30	8.9

Table 8: Result of Correlation between search volume and total box

We only collect 5 item from BAIDU INDEX, but the Correlation between search volume and total box is still perform better than any other attribute like 'actor', 'director'. This attribute can be expected to be a potential feature for linear regression algorithm and SVC algorithm.

9. REFERENCES

[1] Current Situation and Problems of Chinese Film Development. http://rthk9.rthk.hk/mediadigest/20140318_76_123098.html

[2] Quantifying Movie Magic with Google Search https://ssl.gstatic.com/think/docs/quantifying-movie-magic_research-studies.pdf
[3] Predicting Flight Delays with a Random Forest https://lucdemortier.github.io/articles/16/RandomForestsWiMLDS
[4] Scikit-learn - Machine Learning in Python http://scikit-learn.org/stable/index.html
[5] Classification and Regression by randomForest http://www.bios.unc.edu/~dzeng/BIOS740/randomforest.pdf
[6] Mean square error and Root mean error square https://www.vernier.com/til/1014/
[7] Multiple linear regression https://www.statisticssolutions.com/what-is-multiple-linear-regression/
[8] Zero Rule Algorithm http://machinelearningmastery.com/implement-baseline-machine-learning-algorithms-scratch-python/

About the authors:

DENG, Ken is studying the postgraduate program of information technology at Hong Kong University of Science and Technology.

HOU, Jiefeng is studying the postgraduate program of information technology at Hong Kong University of Science and Technology.

LAU, Kim Kwan is studying the postgraduate program of information technology at Hong Kong University of Science and Technology. He currently works as developer in financial software provider company. (www.linkedin.com/in/kim-lau)

WONG, Lok Heung is studying the postgraduate program of information technology at Hong Kong University of Science and Technology. He currently works as developer in Hospital Authority.. (www.linkedin.com/in/lawrence-wong/)